Analysis of Variance

POST READ

Agenda

- Equivalent Non-parametric Test of ANOVA
  - Kruskal-Wallis H Test
- Two way ANOVA
  - Total Variation
  - Treatment Variation
  - Block Variation

Kruskal Wallis H test

Kruskal Wallis H test

- It is the non-parametric equivalent of ANOVA
- Used when two or more independent samples when population is non-normal
- The test hypothesis remains the same as that of ANOVA

Kruskal Wallis H test - Procedure

Great Learning
POWER AHEAD

- Let n_i (i = 1, 2, … ,t) denote the number of observations a treatment
- Combine all t samples so that there are $n = n_1 + n_2 + \ldots + n_t$ units in the sample
- Arrange it in ascending order and rank them (rank 1 is given to the smallest observation, solve for ties)
- The test statistic is given by

sum of ranks of observations in the i^{th} sample

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1)$$

Kruskal Wallis H test

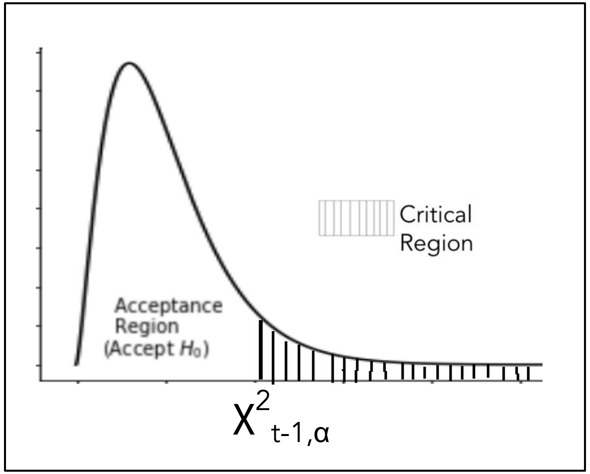
- Note that each $n_i > 5$

- Decision rule: Reject H_0 if $H > \chi^2_{t-1,\alpha}$

  or

  Reject H_0 if p-value $\leq \alpha$

Where, α is the level of significance (l.o.s.)



PLEASE!! NOTE

The python code to conduct kruskal wallis H test is

scipy.stats.kruskal(sample$_1$, sample$_2$,..., sample$_n$)

Kruskal wallis H test

Question:

Ryan is a production manager at an industry manufacturing alloy seals. They have 3 machines - A, B and C. Ryan wants to study whether all the machines have equal efficiency. There no information about the tensile strength (in N/m^2) of the population. Can it be said that the machines are produce the sample tensile strength?

Test at 5% level of significance.

A	B	C
72.5	33.2	24.5
84.8	23.4	34.6
76.6	35.7	
34.7	33.7	
	34.5	

Kruskal wallis H test

Solution:

Ryan is a production manager at an industry manufacturing alloy seals. They have 4 machines - A, B, and C

Let μ_1 be the average tensile strength due to machine A

μ_2 be the average tensile strength due to machine B

μ_3 be the average tensile strength due to machine C

To test, H_0: $\mu_1 = \mu_2 = \mu_3$ Against H_1: At least one μ_i is different (i=1,2,3)

Kruskal wallis H test

Solution:

First combine the sample and rank it.

Here $n_1 = 4$, $n_2 = 5$, $n_3 = 2$

Thus $n = 4+5+2 = 11$

$R_1 = 7+9+10+11 = 37$

$R_2 = 1+3+4+5+8 = 21$

$R_3 = 2+6 = 8$

Sample treatment	Observation	Rank
B	23.4	1
C	24.5	2
B	33.2	3
B	33.7	4
B	34.5	5
C	34.6	6
A	34.7	7
B	35.7	8
A	72.5	9
A	76.6	10
A	84.8	11

Kruskal wallis H test

Solution:

The test statistic is $$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{11(11+1)} [\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3}] - 3(11+1)$$

$$= \frac{12}{11(11+1)} [\frac{37^2}{4} + \frac{21^2}{5} + \frac{8^2}{2}] - 3(11+1)$$

$$= 6.041$$

Kruskal wallis H test

Solution:

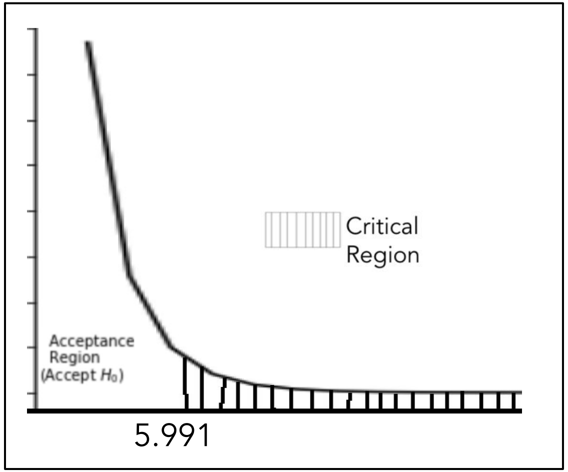
The test statistic H = 6.041

From the chi-square table we have $\chi^2_{t-1,\alpha} = \chi^2_{3-1,0.05} = 5.991$

Since $H > \chi^2_{t-1,\alpha}$ (6.041 > 5.991), we fail to reject H_0

The associate p-value is 0.048
Since p-value < 0.05, we reject H_0.

Thus there is enough evidence to conclude that the tensile strength due to machines is different



Kruskal wallis H test

Python solution:

```
# perform kruskal-wallis H test
test_stat, p_val = stats.kruskal(machine_A, machine_B, machine_C)

# print the test statistic and corresponding p-value
print('Test statistic:', test_stat)
print('p-value:', p_val)
```

```
Test statistic: 6.040909090909089
p-value: 0.04877904103071468
```

As p-value < 0.05, we reject H_0.

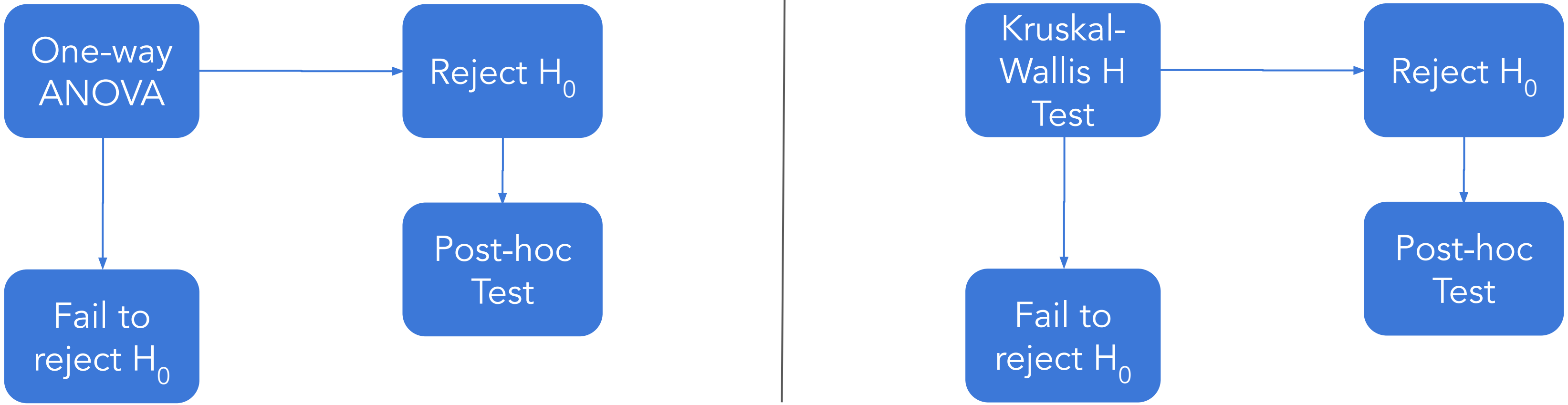
Post-hoc tests

	Parametric test	Non-parametric test
Equal groups	Tukey HSD	Conover test
Unequal groups	Scheffe test	Dunn's multiple Comparison test

Summary

Parametric Test

Non-parametric Test



One - way ANOVA

- Ryan only considered the effect of the machines on the tensile strength
- What if he considers the effect of work shifts used?
- The effect of the quality of material and the effect due to machine can be tested using two way ANOVA

Two way ANOVA

Two way ANOVA

The business is thriving, and Ryan finds it crucial that the production should progress without any breaks. He decides to have three shifts - Morning, Afternoon and Night shift.

Now there are two factors that may affect the tensile strength of the alloy wire.

- The production machine
- The workforce at each shift

Two way ANOVA

- In two way ANOVA the effect due to two nominal variables can be tested over a numeric variable

- Ryan can test for effect of machine on the tensile strength while he tests for effect due to shifts

- Now the effect of machines can be considered as the treatments and the effect due to shifts are the blocks

Blocking

- Previously, Ryan had only one shift that was working. So there was no possible variation that could occur

- Now when he has 3 shifts the and possible variation due to these shifts may arise

- The data becomes heterogeneous and can be made homogeneous

- Blocking is used when the data is not homogeneous

Two way ANOVA

- One of the null hypothesis to be tested is

H_{01}: The averages of all treatments are same.
i.e. $\alpha_1 = \alpha_2 = \ldots = \alpha_n$

Against

H_{11}: At the least one treatment has a different average

- Failing to reject H_{01}, implies that all treatment have the same average

Two way ANOVA

- Another null hypothesis to be tested is

H_{02}: The averages of all blocks are same.
i.e. $\beta_1 = \beta_2 = \ldots = \beta_n$

Against

H_{12}: At the least one block has a different average

- Failing to reject H_{02}, implies that all blocks have the same average

Two way ANOVA

- Suppose Ryan collects data for tensile strength of wires produced by each machine at each shift

- There are 4 treatments (t = 4)

- There are 3 blocks (b = 3)

	A	B	C	D
Morning Shift	68.7	78.7	55.9	84.7
Afternoon Shift	75.4	75.4	56.1	85.3
Night Shift	80.9	80.9	57.3	87.9

Two way ANOVA

- Each treatment has 5 observations (n_i = 5) where i = 1, 2, …, t

- Each block as 4 observations (b_j = 4) where j = 1, 2, …, b

- Total number of observations is given by N

$$N = \sum_i^t \sum_j^b n_{ij}$$

	A	B	C	D
Morning Shift	68.7	78.7	55.9	84.7
Afternoon Shift	75.4	75.4	56.1	85.3
Night Shift	80.9	80.9	57.3	87.9

Two way ANOVA

- Let α_i (i=1, 2, …, t) denote the average strength due to each machine and β_j dnote the average strength due to the shift
- For our example, t = 4 and b = 3
- The test hypothesis can be written as

	A	B	C	D
Morning Shift	68.7	78.7	55.9	84.7
Afternoon Shift	75.4	75.4	56.1	85.3
Night Shift	80.9	80.9	57.3	87.9

H_{01}: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ Against H_{11}: At least α_i is different

H_{02}: $\beta_1 = \beta_2 = \beta_3$ Against H_{12}: At least β_j is different

Two way ANOVA

- In one way ANOVA, the entire population variance is split into two component
  - Variation due to treatments
  - Variation due to blocks
  - Error variation

- Total variation = Treatment Variation + Block Variation + Error Variation

Total variation

- It is the total sum of squares (TSS)
- Let x_{ij} be the observations in the i^{th} treatment and j^{th} block
- $\bar{x}_{..}$ is the grand mean, i.e. the mean of all observations
- The total variation is given by

$$TSS = \sum_i^t \sum_j^b (x_{ij} - \bar{x}_{..})^2$$

Summation over the treatments

Summation over the blocks

Treatment variation

- It is the treatment sum of squares (TrSS)

- Let $x_{i.}$ be the observations in the i^{th} treatment with b observations in each treatment and $\bar{x}_{i.}$ is the mean over i^{th} treatment

- $\bar{x}_{..}$ is the grand mean, i.e. the mean of all observations

- The treatment variation is given by

$$TrSS = \sum_i^t b(\bar{x}_{i.} - \bar{x}_{..})^2$$

Summation over all treatments

Block variation

- It is the block sum of squares (BSS)

- Let $x_{.j}$ be the observations in the j^{th} block with t observations in each block and $\bar{x}_{.j}$ is the mean over j^{th} block

- $\bar{x}_{..}$ is the grand mean, i.e. the mean of all observations

- The block variation is given by

$$BSS = \sum_{j}^{b} t(\bar{x}_{.j} - \bar{x}_{..})^2$$

Summation over all blocks

Error variation

- It is the error sum of squares (ESS)
- Let $x_{i\bullet}$ be the observations in the i^{th} treatment and $\bar{x}_{.j}$ is the mean over j^{th} block
- $\bar{x}_{..}$ is the grand mean, i.e. the mean of all observations
- The error sum of squares is given by

$$ESS = \underbrace{\sum_i^t}_{\text{Summation over all treatments}} \underbrace{\sum_j^b}_{\text{Summation over all blocks}} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2$$

Error sum of squares

During problem solving, the error sum of squares is obtained as:

$$\mathrm{ESS} = \mathrm{TSS} - \mathrm{TrSS} - \mathrm{BSS}$$

Two way ANOVA

- The test statistic for treatments is given by

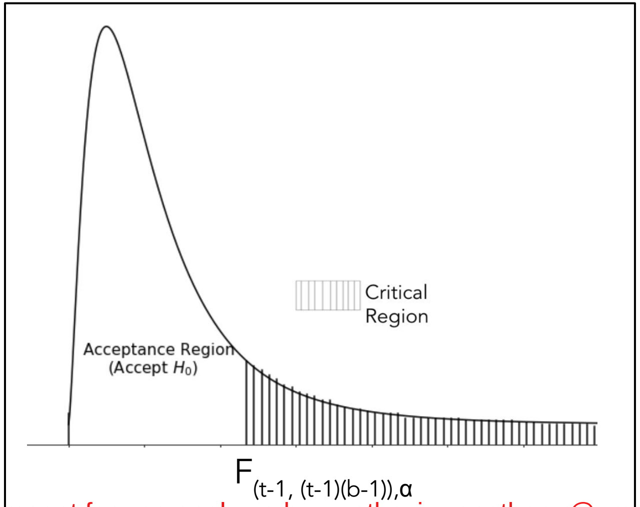
$$\text{F-ratio} = \frac{\frac{TrSS}{df_{Tr}}}{\frac{ESS}{df_e}} = \frac{\underbrace{\text{MTrSS}}_{\text{Mean Treatment Sum of Squares } (s^2_t)}}{\underbrace{\text{MESS}}_{\text{Mean Error Sum of Squares } (s^2_e)}}$$

- Under H_0, the test statistic follows F-distribution with (df_{Tr}, df_e) degrees of freedom

- Decision Rule: If $F_{cal} \geq F_{(t-1,\ (t-1)(b-1)),\alpha}$ then we reject H_0 at α% level of significance

Two way ANOVA

Decision Rule: If $F_{cal} \geq F_{(t-1,\ (t-1)(b-1)),\alpha}$ or p-value $\leq \alpha$, then we reject H_0 at α% level of significance.



Two way ANOVA

- The test statistic for blocks is given by

Mean Block Sum of Squares (s^2_b)

$$\text{F-ratio} = \frac{\frac{BSS}{df_B}}{\frac{ESS}{df_e}} = \frac{\text{MBSS}}{\text{MESS}}$$

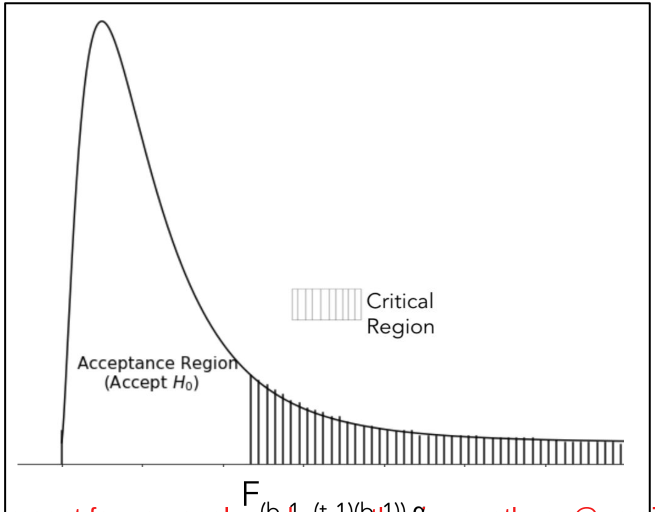
Mean Error Sum of Squares (s^2_e)

- Under H_0, the test statistic follows F-distribution with (df_B, df_e) degrees of freedom

- Decision Rule: If $F_{cal} \geq F_{(b-1,\ (t-1)(b-1)),\alpha}$ then we reject H_0 at α% level of significance

Two way ANOVA

Decision Rule: If $F_{cal} \geq F_{(b-1,\ (t-1)(b-1)),\alpha}$ or p-value $\leq \alpha$, then we reject H_0 at α% level of significance.



Two way ANOVA

The two ANOVA table is given as follows:

Source of variation	Degrees of freedom	Sum of Squares	Mean Sum of Squares	F-ratio
Treatment	t-1	TrSS	s^2_t	$\frac{s_t^2}{s_e^2}$
Block	b-1	BSS	s^2_b	$\frac{s_b^2}{s_e^2}$
Error	(t-1)(b-1)	ESS	s^2_e	
Total	bt-1	TSS		

Two way ANOVA - procedure

1. State the hypothesis to be tested

2. Compute the sum of squares

   a. The total sum of squares, TSS = $\sum_{j=1}^{t}\sum_{i=1}^{b} (x_{ij} - \bar{x}_{..})^2$

   b. The treatment sum of squares TrSS = $\sum_{j=1}^{t} b\,(\bar{x}_{i.} - \bar{x}_{..})^2$

   c. The block sum of squares BSS = $\sum_{i=1}^{b} t\,(\bar{x}_{.j} - \bar{x}_{..})^2$

   d. The Error sum of squares, ESS = TSS - TrSS - BSS

Two way ANOVA - procedure

3. Compute mean sum of squares

   a. s^2_t = Mean treatment sum of squares (MTrSS) = TrSS/(t-1)

   b. s^2_b = Mean block sum of squares (MBSS) = BSS/(b-1)

   c. s^2_e = Mean error sum of squares (MESS) = ESS/[(t-1)(b-1)]

Two way ANOVA - procedure

4. Compute the F-ratios for

   a. Treatment: $\text{F-ratio}_{\text{Tr}} = \frac{\text{MTrSS}}{\text{MESS}} = \frac{s_t^2}{s_e^2}$

   b. Block: $\text{F-ratio}_{\text{B}} = \frac{\text{MBSS}}{\text{MESS}} = \frac{s_b^2}{s_e^2}$

5. Prepare the ANOVA table

6. Write the decision and conclusion accordingly

Two way ANOVA

Question:

The business is thriving, and Ryan finds it crucial that the production should progress without any breaks. He decides to have three shifts - Morning, Afternoon and Night shift.

Ryan collects data of tensile strength (in N/m^2) from all the 4 machines and the 3 shifts as given.

Test at 5% level of significance.

	A	B	C	D
Morning Shift	68.7	78.7	55.9	84.7
Afternoon Shift	75.4	75.4	56.1	85.3
Night Shift	80.9	80.9	57.3	87.9

Two way ANOVA

Solution:

There are 4 machines - A, B, C and D

Let α_1 be the average tensile strength due to machine A
α_2 be the average tensile strength due to machine B
α_3 be the average tensile strength due to machine C
α_4 be the average tensile strength due to machine D

To test, H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ Against H_1: At least one α_i is different (i=1, 2, 3, 4)

Two way ANOVA

Solution:

There are 3 shifts - Morning, Afternoon and Night

Let β_1 be the average tensile strength due to Morning shift
β_2 be the average tensile strength due to Afternoon shift
β_3 be the average tensile strength due to Night shift

To test, H_0: $\beta_1 = \beta_2 = \beta_3$ Against H_1: At least one β_i is different (i=1, 2, 3)

Two way ANOVA

Solution:

The grand mean:

$$\bar{x}_{..} = \frac{68.7+78.7+...+57.3+87.9}{12} = 73.93$$

The total sum of squares

$$TSS = \sum_i^t \sum_j^b (\bar{x}_{ij} - \bar{x}_{..})^2$$

$$= (68.7 - 73.93)^2 + \ldots + (87.9 - 73.93)^2$$

$$= 1511.567 \ (\mathrm{N/m}^2)^2$$

	A	B	C	D
Morning Shift	68.7	78.7	55.9	84.7
Afternoon Shift	75.4	75.4	56.1	85.3
Night Shift	80.9	80.9	57.3	87.9

Two way ANOVA

Solution:

The treatment sum of squares

$$TrSS = \sum_i^t b(\bar{x}_{i.} - \bar{x}_{..})^2$$

$$= 3(75 - 73.93)^2 + \ldots + 3(85.97 - 73.93)^2$$

$$= 1414.647 \ (\mathrm{N/m}^2)^2$$

	A	B	C	D
Morning Shift	68.7	78.7	55.9	84.7
Afternoon Shift	75.4	75.4	56.1	85.3
Night Shift	80.9	80.9	57.3	87.9
$\bar{x}_{i.}$	75	78.33	56.43	85.97

Two way ANOVA

Solution:

The block sum of squares

	A	B	C	D	$\bar{x}_{.j}$
Morning Shift	68.7	78.7	55.9	84.7	72
Afternoon Shift	75.4	75.4	56.1	85.3	73.05
Night Shift	80.9	80.9	57.3	87.9	76.75

$$BSS = \sum_j^b t(\bar{x}_{.j} - \bar{x}_{..})^2$$

$$= 4(72 - 73.93)^2 + 4(73.05 - 73.93)^2 + 4(76.75 - 73.93)^2$$

$$= 49.807 \ \left(\mathrm{N/m}^2\right)^2$$

Two way ANOVA

Solution:

The treatment sum of squares is

$$ESS = \sum_i^t \sum_j^b (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2$$

$$= (68.7 - 75 - 72 - 68.365 + 73.93)^2 + \ldots$$

$$\ldots + (68.7 - 75 - 72 - 68.365 + 73.93)^2$$

$$= 47.1133 \ (\mathrm{N/m^2})^2$$

	A	B	C	D	$\bar{x}_{.j}$
Morning Shift	68.7	78.7	55.9	84.7	72
Afternoon Shift	75.4	75.4	56.1	85.3	73.05
Night Shift	80.9	80.9	57.3	87.9	76.75
$\bar{x}_{i.}$	75	78.33	56.43	85.97	

Or it can be calculated as $\mathrm{ESS} = \mathrm{TSS} - \mathrm{TrSS} - \mathrm{BSS} = 47.1133 \ (\mathrm{N/m^2})^2$

Two way ANOVA

Solution:

Source of variation	Degrees of freedom	Sum of Squares	Mean Sum of Squares	F-ratio
Treatment	t-1 = 4-1 = 3	TrSS = 1414.647	$s_t^2 = \frac{1414.647}{3} = 471.5489$	$\frac{s_t^2}{s_e^2} = 60.0529$
Block	b-1 = 3-1 = 2	BSS = 49.8067	$s_b^2 = \frac{49.8067}{2} = 24.903$	$\frac{s_b^2}{s_e^2} = 3.1715$
Error	(b-1)(t-1) = 6	ESS = 47.1133	$s_e^2 = \frac{47.1133}{6} = 7.852$	
Total	bt-1 = 20 -1 =19	TSS = 1511.567		

Two way ANOVA

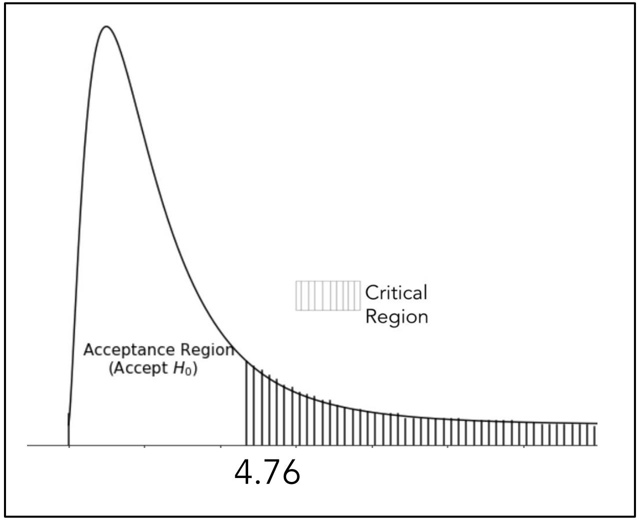
Solution:

For treatments, F-ratio = 60.0529

From the chi-square table we have
$F_{(t-1,(b-1)(t-1)),\alpha} = F_{(3,6),0.05} = 4.76$

Since $F > F_{(3,6),\alpha}$ (6.041 > 4.76), we reject H_0

There is enough evidence to conclude that the tensile strength due to machines is different



Two way ANOVA

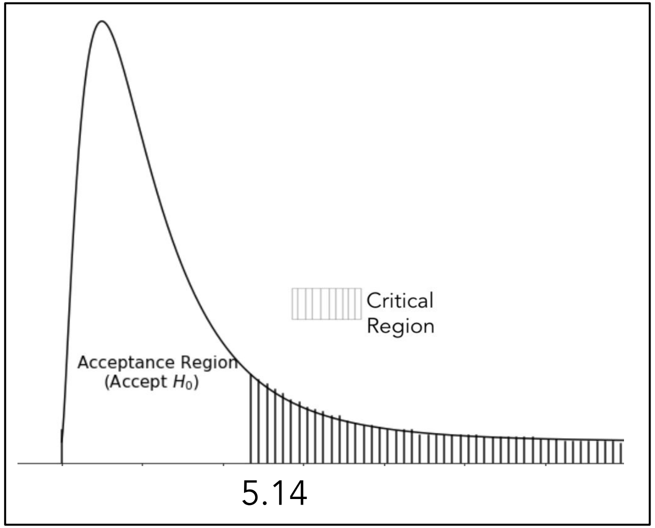
Solution:

For Blocks, F-ratio = 3.17

From the chi-square table we have
$F_{(b-1,(b-1)(t-1)),\alpha} = F_{(2,6),0.05} = 5.14$

Since F < $F_{(2,6),\alpha}$ (3.17 < 5.14), we fail to reject H_0

There is no enough evidence to conclude that the tensile strength due to shifts is different



Two way ANOVA

Python solution:

There is enough evidence to conclude that the tensile strength due to machines is different.

But, There is no enough evidence to conclude that the tensile strength due to shifts is different.

```
# perform two-way ANOVA

# fit an ols model on the dataframe
# use 'Q()' to quote the variable name
# use 'fit()' to fit the linear model
test = ols('Tensile_Strength ~ Q("Machine") + Q("Shift")', df_machine_shift).fit()

# create table for 2-way ANOVA test
# pass the linear model 'test'
# 'typ = 2' performs two-way ANOVA
anova_2 = anova_lm(test, typ = 2)

# print the table
anova_2
```

	sum_sq	df	F	PR(>F)
Q("Machine")	1414.646667	3.0	60.052922	0.000072
Q("Shift")	49.806667	2.0	3.171501	0.114866
Residual	47.113333	6.0	NaN	NaN

Thank You

Appendix

Test statistic for goodness of fit (A.1)

Sometimes the test statistic is given by $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} \quad = \sum_{i=1}^{k} \frac{O_i^2 - 2.O_i.e_i + e_i^2}{e_i}$$

$$= \sum_{i=1}^{k} \frac{O_i^2}{e_i} - 2\sum_{i=1}^{k} \frac{O_i.e_i}{e_i} + \sum_{i=1}^{k} \frac{e_i^2}{e_i}$$

$$= \sum_{i=1}^{k} \frac{O_i^2}{e_i} - 2\sum_{i=1}^{k} O_i + \sum_{i=1}^{k} e_i$$

$$= \sum_{i=1}^{k} \frac{O_i^2}{e_i} - 2N + N$$

...Sum over all the frequencies is the total number of observations

$$= \sum_{i=1}^{k} \frac{O_i^2}{e_i} - N$$

Test statistic for independence of attributes (A.2)

Sometimes the test statistic is given by $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij}-e_{ij})^2}{e_{ij}}$

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij}-e_{ij})^2}{e_{ij}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2 - 2.O_{ij}.e_{ij} + e_{ij}^2}{e_{ij}}$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{e_{ij}} - 2\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}e_{ij}}{e_{ij}} + \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{e_{ij}^2}{e_{ij}}$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{e_{ij}} - 2\sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} + \sum_{i=1}^{r} \sum_{j=1}^{c} e_{ij}$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{e_{ij}} - 2N + N$$

…Sum over all the frequencies is the total number of observations

$$= \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{e_{ij}} - N$$

χ^2 Test for Equality of Variance (A.3)

- The hypothesis to test whether the population variance is equal to a specified value when population mean μ is known

$$H_0 : \sigma^2 = \sigma^2_0 \quad \text{against} \quad H_1 : \sigma^2 \neq \sigma^2_0$$

- It implies

H_0: The population variance is equal to σ_0 (i.e $\sigma^2 = \sigma^2_0$)

against H_1: The population variance is not equal to σ_0 (i.e $\sigma^2 \neq \sigma^2_0$)

Test hypothesis

- Like the test for population mean, it is possible to test for one sided hypothesis

$$H_0 : \sigma^2 \leq \sigma^2_0 \quad \text{against} \quad H_1 : \sigma^2 > \sigma^2_0$$

Or

$$H_0 : \sigma^2 \geq \sigma^2_0 \quad \text{against} \quad H_1 : \sigma^2 < \sigma^2_0$$

- Failing to reject H_0 implies that the population variance is equal to σ_0

Test statistic

- The test statistic is Z given by

μ is known

$$\chi^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma_0^2}$$

- Under H_0, test statistic follows χ^2 with *n* d.f.

Test statistic

- The test statistic is Z given by

μ is known

$$\chi^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma_0^2}$$

- Under H_0, test statistic follows χ^2 with *n* d.f.

μ is unknown

$$\chi^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma_0^2}$$

- Under H_0, test statistic follows χ^2 with *n-1* d.f.

Decision rule

	H_1	Based on critical region		Based on p-value
		μ known	μ unknown	
For two tailed test	$\sigma^2 \neq \sigma^2_0$	Reject H_0 if $\square^2 > \square^2_{n,1-\alpha/2}$ or $\square^2 < \square^2_{n,\alpha/2}$	Reject H_0 if $\square^2 > \square^2_{n-1,1-\alpha/2}$ or $\square^2 < \square^2_{n-1,\alpha/2}$	Reject H_0 if p-value is less than or equal to level of significance
For left tailed test	$\sigma^2 < \sigma^2_0$	Reject H_0 if $\square^2 < \square^2_{n,\alpha}$	Reject H_0 if $\square^2 < \square^2_{n-1,\alpha}$	
For right tailed test	$\sigma^2 > \sigma^2_0$	Reject H_0 if $\square^2 > \square^2_{n,1-\alpha}$	Reject H_0 if $\square^2 > \square^2_{n-1,1-\alpha}$	

Test for equality of variance

Question:

At a dairy, the milk is usually supplied in bottles. The manager does not want to exceed the variance of litres of milk cans to 0.26 l^2. Some the following data gives the litres of milk in the cans.

1.5	1.3	1.5	1.5	1.4	1.7	1.6	1.2

Write the null and the alternative hypothesis. Test whether the variance of litres of milk cans is more than 0.26 l^2 at 5% los.

Test for equality of variance

Solution:

Let X: Litres of milk in the can

Here $\sigma^2_0 = 0.26\ l^2$ and $n = 8$

To test, $H_0: \sigma^2 \leq 0.26$ Against $H_1: \sigma^2 > 0.26$

The population mean is unknown. So we estimate it from the sample

$\bar{x} = 1.4625$

Test for equality of variance

Solution:

The test statistic is computed as $\chi^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma_0^2}$

$$= \frac{(1.5 - 1.4625)^2 + \ldots + (1.2 - 1.4625)^2}{0.26}$$

$$= \frac{0.17875}{0.26}$$

$$= 0.6875$$

Test for equality of variance

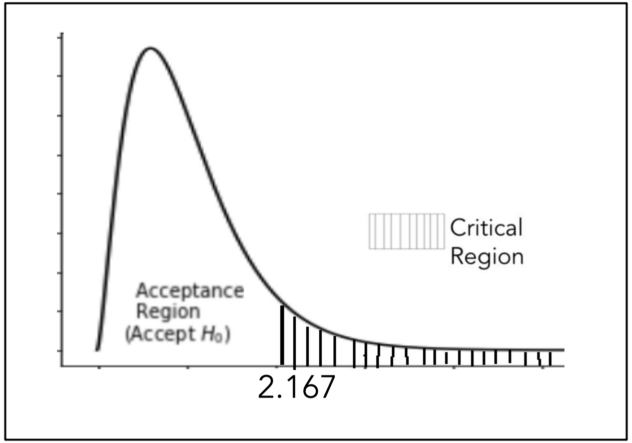
Solution:

The test statistic is 0.6875.

It follows χ^2 with *n-1* d.f. That is n-1 = 8-1 =7

Thus $\chi^2_{n-1,1-\alpha} = \chi^2_{7,0.95} = 2.167$

Since $\square^2 < \square^2_{n,1-\alpha}$, we fail to reject H_0 i.e. we accept H_0



Thus the variance of litres of milk cans is less than 0.26 l^2.

Test for equality of variance

Python solution:

```
# given data
milk_lit = [1.5, 1.3, 1.5, 1.5, 1.4, 1.7, 1.6, 1.2]

# hypothesized variance
sig_2 = 0.26

# sample size
n = len(milk_lit)

# the population mean is unknown, use sample mean
samp_mean = np.mean(milk_lit)

# calculate the test statistic
chi_test = (np.sum((milk_lit - samp_mean)**2)) / sig_2

# calculate P(X > chi_test)
# pass the test statistic value to 'x'
# pass the degrees of freedom to 'df'
p_val = stats.chi2.sf(x = chi_test, df = 7)

# print the p-value
print('p-value:', p_val)
```

```
p-value: 0.9984299481373 78
```

As p-value > 0.05, we fail to reject H_0.

Non-parametric post hoc test (A.4)

- The post hoc test performed after non-parametric Kruskal Wallis H test is conover test
- It test is based on studentized range
- The hypothesis remains the same as that of a parametric post-hoc test

Non-parametric post hoc test

Thus the test hypothesis for are our example are

$H_{01}: \mu_{machine\_A} = \mu_{machine\_B}$ Against $H_{11}: \mu_{machine\_A} \neq \mu_{machine\_B}$

$H_{02}: \mu_{machine\_A} = \mu_{machine\_C}$ Against $H_{12}: \mu_{machine\_A} \neq \mu_{machine\_C}$

$H_{03}: \mu_{machine\_B} = \mu_{machine\_C}$ Against $H_{13}: \mu_{machine\_A} \neq \mu_{machine\_C}$

Machine A → Machine B

Machine A → Machine C

Machine B → Machine C

Non-parametric post hoc test

- The post hoc test performed after non-parametric Kruskal Wallis H test is Conover test

- The python code to conduct non-parametric post hoc test is

```
# perform the conover test
# pass the dataframe to the parameter, 'a'
# pass the column with numeric data to the parameter, 'val_col'
# pass the column with categoric data to the parameter, 'group_col'
scikit_posthocs.posthoc_conover(a = df_machine, val_col = 'strength', group_col = 'machine')
```

Non-parametric post hoc test

The output is as follows:

	machine_A	machine_B	machine_C
machine_A	1.000000	0.012120	0.031704
machine_B	0.012120	1.000000	0.920917
machine_C	0.031704	0.920917	1.000000

Reject H_0 if p-value is less than α

If $\alpha = 0.05$, it can been seen that there is statistical difference between pairs of machines (A,B) and (A,C).